PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : H04L 12/56

A1

(11) International Publication Number: WO 00/24164

(43) International Publication Date: 27 April 2000 (27.04.00)

(21) International Application Number: PCT/US99/24145

(22) International Filing Date: 15 October 1999 (15.10.99)

(30) Priority Data:
09/176,061 20 October 1998 (20.10.98) US

(71) Applicant (*for all designated States except US*): AMADON & ASSOCIATES, INC. [US/US]; 1017 East Blaine Street, Seattle, WA 98102 (US).

(72) Inventors; and

(75) Inventors/Applicants (*for US only*): ZIKAN, Karel [US/US]; 4640 Sunnyside Avenue North, Seattle, WA 98103 (US). SOWIZRAL, Henry, Adam [US/US]; 16 East Portula Avenue, Los Altos, CA 94022 (US).

(74) Agents: RONDEAU, George, C., Jr. et al.; Seed and Berry LLP, 6300 Columbia Center, 701 Fifth Avenue, Seattle, WA 98104–7092 (US).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published

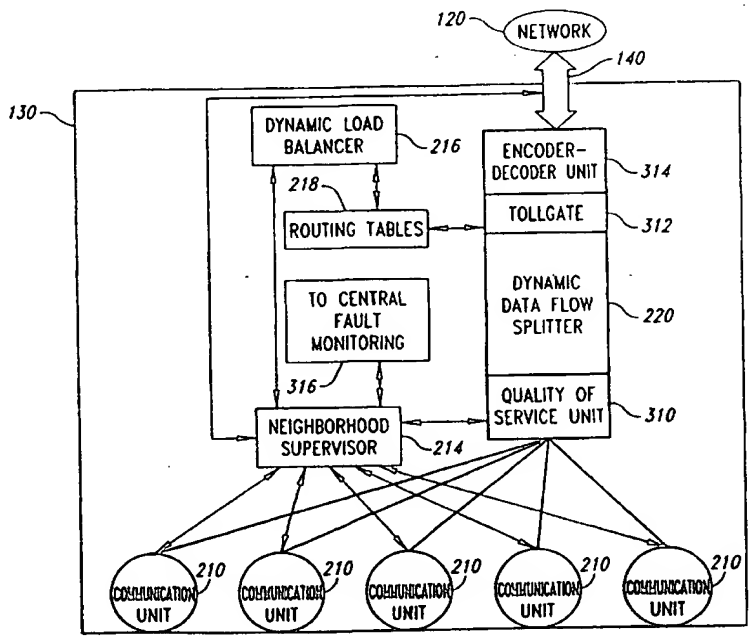
With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: METHOD AND APPARATUS FOR NETWORK CONTROL

(57) Abstract

In a data communication system having network traffic with flows, a network traffic director system using router modules. The router modules direct traffic based on optimizing a merit function or penalty function to reduce costs of congestion for stochastically changing demands and flows in the data communication system. The router modules exchange values with neighboring router modules. Based on the exchanged values and values local to a router module, flow conditions are checked and if necessary the local values are adjusted until the flow conditions are satisfied or a time period expires. Adjustments are associated with optimizing a merit function or penalty function. Based on the adjusted values, the router module adjusts parameters to be used to direct packets of the network traffic flows to other router modules or other destinations within the data communication system. An aggregation scheme is used for reducing the number of values stored in a single router module.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

METHOD AND APPARATUS FOR NETWORK CONTROL

TECHNICAL FIELD

The present invention is related generally to data communication systems and, in particular, to network control of data communication systems.

BACKGROUND OF THE INVENTION

Data communication systems generally consist of multiple networks that are linked together with bridges, switches, routers, or other devices. These linking devices direct data packets along various paths from an origination to a destination. Since a data communication system tends to be large with many networks and interconnecting links, typically a data packet can take many paths from its origination to its destination. A goal of a network traffic director system is to direct the journey of each data packet from its origination to its destination so that the capacity of the network is efficiently utilized.

Traditional network traffic director systems utilize a plurality of network routers based on Internet Protocol (IP). The IP routing approaches are based on a Shortest Path (SP) method in which packets are sent to their destinations along the shortest paths in the network. There are various methods to determine shortest paths among the IP routing approaches. However, these approaches all share a common assumption that if each packet travels along the shortest path, then the overall workload required for the routing will remain as small as possible.

The assumption of traditional routing systems would hold true if the overall performance of a routing system could be adequately measured solely on the basis of the total number of decisions that routers need to perform. However, other factors are involved in the overall performance of a routing system. One of these factors is the amount of congestion in a data communication system. Some theoretical proposals exist which attempt to address the congestion of network traffic director systems. These approaches, however, are impractical in providing adequate solutions to the congestion problem. The end result of some of these approaches produces, on

average, flows that are little different than the shortest path methods of IP routing taking congestion into account. Another theoretical approach to congestion is intended for environments having predefined aspects, such as loading and network topology. However, a data communication system is seldom sufficiently predictable for these predefined approaches.

Other theoretical approaches attempt to address an aspect of the stochastic (random) nature of flows in a data communication system by allowing for randomness in the arrival times of individual data packets into a linking device. These approaches, however, are still formulated in a deterministic manner by requiring that demand or load rates and flow rates in other parts of the communication system be known in advance of their actual occurrence.

These prior art deterministic approaches are not ideally suited for data communication system environments in which demands or loads and flows in other portions of the data communication system have stochastic distributions due to changing demands or loads of individual users or due to any multiplexing involved. This means that not only are the arrival times of individual packets random, but also the rates of the demands or loads and flows in other portions of the data communication system are random and not known before their actual occurrence.

An effective optimization method must not only minimize congestion in the data communication system but it must be robust enough to adapt to sudden changes in conditions of the data communication system such as packet arrival, loading on the system, and system topology without predefined scenarios of when or how these changes will occur.

In data communication systems, loading changes quickly and dramatically due to such influences as changing demands or loads of individual users and multiplexing methods such as time division multiplexing (TDM) where time slots of system use are divided among individual users. With TDM, a data communication system constantly changes the particular users on the system at any given moment, thus loading changes rapidly. Network topology can also have rapid stochastic changes due to situations such as equipment failures or weather conditions.

Also, whereas the prior art systems and methods focus on a single indicator, such as shortest path or average delay, as a basis to optimize data flow, there are many additional factors. These include average queue length at various data communication system linking devices, variance or standard deviation in individual delays of data packets traveling from an origination to a destination, and average and variance of the individual utilized capacity of links and linking devices in a data communication system. Thus, a system and method is greatly desired which optimizes packet flow in a data communication system based on numerous factors while being robust enough to effectively adapt to unexpected events.

SUMMARY OF THE INVENTION

The present invention is directed to a network traffic director system in a data communication system having network traffic with randomly distributed demands and flows. In one aspect of the present invention the network traffic director system includes router modules configured to direct data packets in the data communication system, each router module being a home router module including a neighborhood supervisor. The neighborhood supervisor is configured to send home potentials of the home router module to neighborhood supervisors of neighboring router modules and to receive neighbor potentials of the neighboring router modules from neighborhood supervisors of neighboring router modules. The home router module further includes a dynamic load balancer configured to determine flows based on the home and neighbor potentials. The dynamic load balancer adjusts the home potentials if first conditions including flow conditions are not met. The dynamic load balancer also updates routing tables if second conditions based on the adjusted home potentials are met. The home router module also includes a dynamic data flow splitter configured to receive data packets from networks and router modules. The dynamic data flow splitter selects a portion of the data communication system for each data packet received based on the updated routing tables. Each received data packet is transmitted to the portion of the data communication system selected by the dynamic data flow splitter for the received data packet.

Another aspect of the invention includes the home potentials of each home router module being associated with a pair of nodes of the data communication system including an origination node and a destination node. The home potentials of each router module are also associated with a quality of service level. The neighborhood supervisor is configured to send the home potentials to neighborhood supervisors of neighboring router modules when the flow conditions are not met. The flow conditions include conservation of flows. The dynamic load balancer is configured to determine flows by using a response function associated with optimizing at least one of a penalty function or a merit function for stochastic demands or flows of the data communication system. The dynamic data flow splitter is configured to select the portion of the data communication system for each packet received further based on either a Markov protocol, a Routing Wheel protocol, previously selected portions of the data communication when an address for a received data packet is a same type as an address stored in the dynamic data flow splitter, or on a quality of service level associated with a data packet.

In a further aspect of the invention the neighborhood potentials are stored according to an aggregation scheme. The merit or penalty functions are at least locally approximated by a quadratic function of the flows. The ideal data flows are further based on resistance of arcs associated with the router module wherein each resistance of an arc is configured to be a function of either arc capacity or a function of flows of the arc. The neighbor router modules are configured to be associated with the home director module through at least one of a political, organizational, topological, topical, or geographical relation. In a further aspect each router module is configured to adjust home potentials to an equilibrium point for a difference function associated with flows of arcs associated with the router module. The difference function is associated with differences between home potentials and neighbor potentials. The merit function or penalty function involves costs of congestion including at least one of traffic delays, high latency, diminished throughput, lost packets or unresponsiveness to sudden changes in topology or loading.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a topology diagram representative of a network traffic director system of the present invention.

Figure 2 is a block diagram showing interconnection detail of router modules of the network traffic director system of Figure 1.

Figure 3 is a block diagram showing internal detail of a router module of Figure 2.

Figure 4 is a topology diagram showing internal detail of a router module of Figure 2.

Figure 5 is a relational diagram of an aggregation scheme used in the network traffic director system of Figure 1.

Figure 6 is a topology diagram showing internal detail of an arc of a router module of Figure 2.

Figure 7 is a flowchart of a method used by the router modules of Figure 2.

DETAILED DESCRIPTION OF THE INVENTION

Large communication data systems are comprised of a collection of networks linked together by linking devices, such as routers, bridges, and switches known herein as router modules or director modules. The depicted embodiment in Figure 1 has a network traffic director system 110 in a data communication system which links a collection of networks 120 via router modules 130 and associated links 140 and links 150. A link 140 is a data communication path for data packets to travel between a network 120 and a router module 130. A link 150 is a data communication path for data packets to travel between two router modules 130. Network traffic is made up of individual data packets which are sent between the networks 120 via the router modules 130. The network traffic director system 110 as embodied in the router module 130 allows more efficient utilization of data communication systems over prior art linking devices. The router module 130 is configured to adapt to sudden changes in the data communication network, such as randomly and rapidly changing service

demand or load rates or topology, while optimizing utilization of the data communication system based on numerous factors, including traffic delays, queue lengths, variance in individual delays, average and variance of utilized capacity of links and linking devices, number of lost packets, latencies, and throughput of linking devices.

Unlike prior art methods, for the depicted embodiment, the optimization methods used account for the numerous factors influencing utilization of a data communication system by formulating a penalty function (which accounts for negative effects, such as lost packets, increased delays, high latency, diminished throughput, among other things) or its opposite merit function (which accounts for beneficial effects). The merit and penalty functions are formulated for optimization and determination of equilibrium points given a rapidly changing, stochastic environment of random demands upon a data communication system, random flows through a data communication system and random changes in topology of a data communication system.

This stochastic approach is far different from prior art methods, which rely on predictable, non-stochastic environments and utilize deterministic formulations. The systems and methods used by the depicted embodiment of the network traffic director system 110 are designed to be effectively implemented in a data communication system without having extraordinary processing requirements, resource needs or other demands upon the overall data communication system. These and other advantages will become apparent in the following detailed description. In the following description, numerous specific details are provided such as methods and systems used to collect information from the router modules 130, to determine whether conditions are satisfied, and to adjust values related to the router modules 130 accordingly. One skilled in the relevant art, however, will recognize that the invention can be practiced without one or more of the specific details, or with other devices and methods. In other instances, well known structures or operations are not shown or described in detail to avoid obscuring the description of the embodiments.

In order to connect the networks 120 of the data communication system together, the router modules 130 are connected together via communication units 210 as shown in Figure 2. The communication units 210 are linked together by data communication pathways 212. The communication units 210 in one embodiment are Ethernet network cards and the data communication pathways 212 are Ethernet cables. Ethernet is a term well known in the art referring to collision sense multiple access/collision detection techniques as described under the IEEE 802 series specifications. In other embodiments of the present invention, the communication units 210 and pathways 212 are other types of point-to-point data communication links.

Some of the communication units 210 shown in Figure 2 are not shown to be connected via any data communication pathway 212. These communication units 210 can be connected to other router modules 130 which are not shown in Figure 2. The number of communication units 210 shown for each individual router module 130 is not intended to limit the depicted embodiment since some router modules have more communication units 210 than shown in Figure 2 if necessary to link additional router modules.

Each router module 130a-d, as shown in Figure 2, is connected to a data network 120a-d. However, the number of data networks 120 connected to any one router module 130 varies, and some router modules 130 are only used to interconnect other router modules and are not directly connected to any network 120. In other instances, router modules 130 can be directly connected to a plurality of networks 120. Each router module 130 has a neighborhood supervisor unit 214 that includes a values exchanger, a dynamic load balancer unit 216, a routing table unit 218 with a plurality of routing tables, and a dynamic data flow splitter unit 220.

The neighborhood supervisor unit 214 of a home router module 130 exchanges information with router modules that are neighbors to the home router module. These neighboring router modules 130 are directly connected to the home router module. For instance, in Figure 2 the neighbors of home router module 130a include router modules 130b, 130c, and 130d because they are directly connected to router module 130a via communication pathways 212 from one communication unit

210 of a neighboring router module 130b-d to a communication unit 210 of the home router module 130a.

The neighborhood supervisor unit 214 also monitors for malfunctions and link failures in the data communication system. Along with monitoring, the neighborhood supervisor unit 214 updates data in the dynamic load balancer 216 of the home router module 130 to be consistent with the current physical state of the data communication system. The neighborhood supervisor unit 214 also reports detected network failures to a central fault reporting device located in the data communication system through a central fault monitoring function.

The dynamic load balancer unit 216 uses information collected by the neighborhood supervisor unit 214 of the home router module 130 from the neighboring router modules 130. The dynamic load balancer unit 216 adjusts the routing tables of the routing table unit 218 based upon the information collected in order to optimize overall utilization of the data communication system served by the network traffic director system 110. The methods used by the dynamic load balancer unit 216 will be discussed further below. The routing table unit 218 provides information to the dynamic data flow splitter unit 220 of the home router module 130 in order that the data packets received by the communication unit 210 of the home router module 130 are directed to destination networks 120 or router modules 130 in a manner that preserves the integrity of the mathematical models used by the dynamic load balancer unit 216.

The dynamic data flow splitter unit 220 performs the actual routing operation of data packets received by the home router module 130 to direct the received data packets to other networks 120 or router modules. The division of labor between the routing table unit 218 and the dynamic flow splitter unit 220 can be varied. In one embodiment, the routing table unit 218 prepares the routing data in a suitable form for the dynamic data flow splitter unit 220. The routing table unit 218 then stores this routing data until the dynamic data flow splitter unit 220 requires the data. The dynamic data flow splitter unit 220 then looks up the routing data in the routing table unit 218 to determine how a particular data packet is to be routed. The dynamic data flow splitter unit 220 in various embodiments of the present invention use different

methods to determine how a received data packet should be routed. Depending on which method is used, a received data packet will be sent by either the communication unit 210 or the dynamic data flow splitter 220 of the router module 130 to different router modules 130 through one of the communication units 210 or to a directly linked network 120. Four different methods of determination, also called splitting protocols, are typically chosen.

Markovian splitting is the simplest of these routing protocols to implement in software. In the Markovian protocol, the routing table unit 218 transforms the desired flows into equivalent routing fractions. Only the flows out of the router module 130 are relevant. For each destination, if f_i is the desired flow to the neighbor i for all the neighbors j of the router module 130, then the fraction of the flow to this neighbor is $p_i = f_i / \sum_j f_j$. In turn, the dynamic data flow splitter unit 220 views the fraction as a probability of a Markov process. When routing a data packet, the dynamic data flow splitter unit 220 looks up the probabilities corresponding to the packet destination, $p_1, p_2, \ldots, p_K$. The flow splitter then "casts a random die," with an associated probability p_i of the outcome i, and sends the data packet to the lucky winner.

The easiest way to cast the random die is to generate a uniformly random number between zero and one and determine where this number falls in the cumulative probability sequence $\{P_i\}$ given by $P_0 = 0$, $P_i = P_{i-1} + p_i$. The random numbers can be generated "on the fly" or they can be stored *a priori*.

The second routing protocol, called a "Routing Wheel," reduces the randomness of the Markovian protocol and is easy to implement in hardware. Given the desired probabilities $p_1, p_2, \ldots p_K$, for flows to 1 through K neighboring router modules 130 of a given router module, the routing table unit 218 of the given router module generates N (*e.g.*, 100) labels. The labels identify individual router modules 130 and are assigned among the neighboring router modules 130 according to the probability associated with each neighboring router module. For instance, the neighboring router module 130 "i" having probability p_i is assigned Np_i number of labels identifying that particular neighboring router module. The routing table unit 218

then arranges the labels into a conceptual distribution resembling a physical circle where all label types are distributed fairly evenly around the conceptual circle. The routing table unit 218 then sets a conceptual pointer to a location on the circle. When the dynamic data flow splitter unit 218 services a data packet, it looks up the currently selected label on the circle. The dynamic data flow splitter unit 220 then sends the packet to the neighboring router module 130 indicated by the current label and then advances the pointer to the next label on the circle. The process never runs out of labels since when the pointer is at the label in position N, the pointer next advances to the label in position 1. In further embodiments, a routing wheel is generated for each origination/destination pair and quality of service level combination (*OD/QoS*). *OD QoS* combinations are discussed in detail below.

Some higher level IP protocols are designed in such a way that makes the arrival of the data packets in an orderly sequential fashion at the final destination desirable. These protocols can handle fallouts and out-of-sequence arrivals, but they become increasingly inefficient as the number of out-of-sequence arrivals grows. In order to increase the efficiency of such transfer protocols, the following modification of flow splitting is used. This modification is incorporated into the third routing protocol and works with either Markovian splitting or Routing Wheel splitting.

For the third routing protocol, the granularity of splitting is changed for the applications running the higher level protocols mentioned above. Such protocols can usually be recognized from a device number in a packet header. The dynamic flow splitter unit 220 is endowed with short-term memory. When one of the "suspect" device numbers appears in the header of a packet, the dynamic flow splitter unit 220 remembers the origin, the destination, the device number of this packet, as well as the label of the neighboring router module 130 to which the packet was sent. All subsequent packets that have the same origin, destination, and device number that arrive while the memory persists are then sent to the same neighboring router module 130 as the original packet. In alternative embodiments, the short-term memory is optionally refreshed by the later arrivals.

The fourth routing protocol used with the dynamic data flow splitter unit 220 relates to low latency and quality of service. Latency of a system is the time required to send and receive information. Some applications, such as telephony, video-telephony, or real-time interactive games or simulations require low latency levels. The latency requirements of other applications such as transmissions of email are not as stringent. The ordinary method for lowering latency for urgent data packets is to give these packets a routing priority while in routing queues as discussed below.

Achieving additional speedups of urgent data packets is realized by making the dynamic data flow splitter unit 220 send more of the urgent packets over faster links and paths. The dynamic data flow splitter unit 220 then compensates for this by sending more of the ordinary packets over slower links. The compensation can be adjusted to leave the overall performance of the routing system unchanged. In one embodiment, identification of the faster links by the network traffic director system 110 in the data communication system is readily accomplished because a high correlation exists between the highly used links and the faster links. This condition is used to satisfy latency and quality of service requirements by directing the urgent packets toward the most frequented path choices for a given address group and by directing more of the less urgent packets away from the most frequented path choices.

The router module 130 also has a quality of service unit 310, a toll gate unit 312, an encoder-decoder unit 314, and a central fault monitoring unit 316 as shown in Figure 3. The quality of service unit 310 is concerned primarily with latency regarding transmission of data packets in the data communication system. The quality of service unit 310 is used to establish separate queues for applications with different latency and bandwidth needs. The queues are then served by appropriate frequencies of service and processing rates. Also, billing rates are dependent upon the level of quality of service. The toll gate unit 312 collects the billing statistics used to bill customers of the data communication system. These statistics primarily focus on billing. However, they can be used to determine bandwidth demands and optimize routing network topology.

To ensure privacy of communication, the data packets sent across network links that are vulnerable to eavesdropping are encoded for transmission and decoded at the destination router module. The encoder-decoder unit 314 performs the encoding and decoding operations. This link security is not directly related to the routing operations of the router module 130. Sufficient bandwidth for the encoder-decoder unit 314 is required so that transmission speed of the overall router module 130 is not impeded.

As described above, the dynamic load balancer unit 216 uses information from the neighborhood supervisor unit 214 to determine parameters that the routing table unit 218 then uses to prepare routing table data. A goal of the dynamic load balancer unit 216 is to optimize traffic flows in the data communication system of which the network traffic director system 110 is part. A method used by the dynamic load balancer 216 of the depicted embodiment used to accomplish this goal is described as follows.

As discussed above, each link 140 between one router module 130 and one network 120 and each link 150 between two router modules 130 is a data communication path that allows for two-way traffic of data packets. Thus, each of links 140 and 150 of Figure 1 allows data packet traffic to travel in two directions either to or from a particular network 120 or router module 130.

To assist in describing the systems and methods below, each direction that a packet can travel along a link 140 or 150 is referred to as an "arc." The term "arc" does not refer to any particular physical part of a link 140 or 150 because the links allow packets to travel in either one of two directions. Rather, the term "arc" is conceptual since it refers to only one direction of travel in a link 140 or 150. Since links 140 and 150 each allow for two directions that packets can travel, each link 140 and 150 is assigned two conceptual arcs, one arc for each of the two directions of travel allowed for within each link. Thus, each line of Figure 1 which represents a link 140 or 150 could be replaced by two lines representing two arcs which conceptually indicate the two directions of travel that are possible for packets to take on the link. The arcs of

links 140 and 150 are illustrated in Figure 4 where each link has two arrows representing arcs for the two directions that packets can travel for any one link.

The depicted embodiment is not limited to links between networks 120 and router modules 130 but could include links between other network devices or networks of any size and topology. For example, links between routers and individual computers or terminals and links between large networks such as the world-wide Internet and a country sized corporate networks and small office local area networks are included in the depicted embodiment.

Within router modules 130 there is an *In* node 420 and an *Out* node 430 as shown in Figure 4. Flow of the internal arcs 440 originate from the *In* nodes 420 and terminate at the *Out* nodes 430 of each router module 130. The internal arcs 440 allow for additional parameters to be used in optimizing the data communication system. To furnish additional parameters, each *In* node 420 and *Out* node 430 is treated similarly as how the router modules 130 are treated in the various formulations for the depicted embodiments discussed below.

For typical flow conditions in a data communication system, an overall flow in a particular arc typically is a conglomeration of one or more separate flows. Each separate flow is characterized by its particular combination of flow origination and flow destination expressed as its origination/destination (*OD*) pair. The origination and destination are typically expressed at the network 120 or router module 130 level, so it is typically not expressed at a level specifying particular devices attached to a network. In some embodiments, the separate flows are characterized only by *OD* pairs, so for these embodiments, the separate flows for a particular arc have unique *OD* pairs. Summing over these unique *OD* pairs gives the overall flow f_{arc} for these arcs.

In the depicted embodiment, the separate flows are also characterized by different quality of service (*QoS*) levels. In this embodiment, at least occasionally, a particular arc will have more than one separate flow having the same *OD* pair, but having different *QoS* levels. To determine the overall flow f_{arc} for these types of arcs, the separate flows would be summed over both the *OD* pairs and *QoS* levels, giving unique *OD/QoS* combinations. Characterization of the separate flows of an arc is not

limited in the present invention to only *OD/QoS* combinations, but also incorporates other factors associated with traffic flow in a data communication system. For instance, in another embodiment, characterization of separate flows in an arc includes a device number of an origination or destination device attached to a network 120.

The overall flow of an arc is also constrained by an upper bound and a lower bound. The upper bound is typically proportional to the processing capacity of the arc for the depicted embodiment (*e.g.*, 0.7 or 0.8 of the processing capacity of the arc). Each individual flow can also be constrained by its individual upper bound and individual lower bound. Again, the most common lower bound is zero for the depicted embodiment.

The method to optimize traffic flow used by the dynamic load balancer unit 216 of the depicted embodiment defines an ideal theoretical state for data packet traffic in the data communication system as the case where there is zero flow throughout the data communication system while all demands or loads by users upon the data communication system are satisfied. A solution for data flows in the data communication system is sought where the solution is a point in a solution space that has a minimum distance to the theoretical ideal state. The solution is subject to certain physical constraints of a data communication system and is found through convex duality theory or convex optimization theory.

In the depicted embodiment, the solution for data flows in a data communication system is uniquely formulated to optimize a variety of network objectives and conditions compared with prior art formulations that focus on one known data rate objective or condition subject to conservation of data flows. Data flows are conserved when the amount of flow coming into a node (typically a router module 130) equals the amount of flow leaving the node. Any data either produced or destroyed at a node is factored into data flow conservation.

The solution for data flows also optimizes the following uniquely formulated expression $E_{\alpha,\beta}(f)$ involving a substantially quadratic function of data flows in a data communication system:

$$E_{\alpha,\beta}(f) = \alpha \sum_{j} \sum_{ab} \left[(r_{j,ab}/2) * (g_{j,ab} - f_{j,ab})^2 \right] +$$

$$\beta \sum_{ab} \left[(r_{ab}/2) * \left(g_{ab} - \sum_{j} f_{j,ab} \right)^2 \right]$$

where the subscript "j" indicates a particular *OD/QoS* combination and the subscript "ab" indicates a particular arc "ab". The term $f_{j,ab}$ is for flow associated with a particular *OD/QoS* combination "j" and a particular arc "ab". The terms $r_{j,ab}$ and $g_{j,ab}$ are resistance and goal parameters, respectively, for the particular *OD/QoS* combination "j" and arc "ab". The multiplication factors α and β are used to tailor emphasis between the two component expressions for $E_{\alpha,\beta}(f)$. Since only the relative magnitude of α and β are important, we implicitly assume the quantities are scaled to satisfy $\alpha + \beta = 1$.

In the form above, the expression $E_{\alpha,\beta}(f)$ is a penalty function which represents undesirable influences and results affecting communication flow in a data communication system as have been described. Optimization of the expression $E_{\alpha,\beta}(f)$ as a penalty function would result in minimization of $E_{\alpha,\beta}(f)$. If the expression $E_{\alpha,\beta}(f)$ was negated it would take on the form of the merit function. Optimization of a merit function would result in a maximization of the merit function. Other formulations that are equivalent to optimization of the penalty and merit functions involve determining a point of equilibrium. As known in the art, once an optimization formulation is described, an equivalent formulation for a point of equilibrium is determined by using mathematical techniques known in the art.

In a preferred embodiment, $r_{j,ab} = r_{ab} = 1/c_{ab}$; $g_{j,ab} = g_{ab} = 0$; and α and β are of the same magnitude. Then the solution of the optimization problem is the same as the solution of the competitive equilibrium problem using only the latter part of the equation, that is

$$E_{\alpha,\beta}(f) = \sum_{ab} \left[(1/2\,c_{ab}) * \left(\sum_{j} f_{j,ab} \right)^2 \right]$$

over the non-supply/demand arcs.

Other embodiments also include minimization of other penalty functions and maximization of other merit functions. Such penalty and merit functions are related to the depicted embodiment in that portions of these other penalty and merit functions are at least locally approximated by the expression $E_{\alpha,\beta}(f)$ above. These penalty and merit functions all are advancements compared with the prior art, since they are used to optimize flows in a stochastic environment of a data communication system.

The expression $E_{\alpha,\beta}(f)$ incorporates factors associated with individual *OD/QoS* combinations for each arc "*ab*" over all the arcs in a data communication system. Each individual *OD/QoS* combination "*j*" of an arc "*ab*" has a particular resistance parameter, $r_{j,ab}$ related to any hindrances to flow and a goal parameter $g_{j,ab}$ related to the current demand or load by users.

For the depicted embodiment, the resistance parameter for an individual arc "*ab*" and *OD/QoS* combination "*j*" is also uniquely formulated, being proportional to the reciprocal of the overall data flow capacity, $C_{j,ab}$ of the individual arc "*ab*" and *OD/QoS* combination "*j*". Thus, $r_{j,ab} \propto 1/C_{j,ab}$, where $C_{j,ab}$ is the maximum amount of data flow which the individual arc "*ab*" is capable of supporting for *OD/QoS* combination "*j*". For instance, $C_{ab} = C_{j,ab}$ *and* $r_{ab} = r_{j,ab} \propto 1/C_{j,ab}$ for all *j*. In component flows, most commonly $g_{ab} = \sum_j g_{j,ab}$. The resistance parameter is also used to designate that an arc for an *OD/QoS* combination is not functional by setting $r_{j,ab}$ to a relatively large number. The solution to the optimization of the uniquely formulated expression $E_{\alpha,\beta}(f)$ over all the component flows $f_{j,ab}$ of each *OD/QoS* combination "*j*" for each arc "*ab*" over all arcs and *OD/QoS* combinations of a data communication system results in solutions of flow $f_{j,ab}$ for each *OD/QoS* combination "*j*" for each arc "*ab*" in the data communication system.

Methods that could be used to solve the optimization of the uniquely formulated expression $E_{\alpha,\beta}(f)$ include methods that have been used to solve other optimization formulations such as a projected gradient method, and a hill climbing method. The depicted embodiment uses a unique distributed optimization method incorporated in the dynamic load balancer unit 216 for optimizing the expression

$E_{\alpha,\beta}(f)$. This method is uniquely formulated for stochastic network flow that is uncertain, rapidly changing and random and topology that can stochastically change. The distributed optimization method is generally easier to implement and converges more quickly to a solution compared with prior art methods that have been used on other formulations. The distributed optimization method also incorporates conservation of flows and unique formulations of g_{jsah} and r_{jsah}.

As part of this distributed optimization method, the depicted embodiment defines a set of quantities at each router module 130 that are related to flows. Each quantity is referred to as a "potential." The set of potentials for a particular router module 130 has one potential for each possible *OD/QoS* combination known to the particular router module 130. Possible *OD* pairs are determined by the various combinations of a data packet origination and a data packet destination. These combinations would not typically include combinations where both the origination and destination are the same. Typically, the data packet origination and destinations would be networks 120 and router modules 130, but could include other network devices at a more discrete level as discussed above, such as individual computers or terminals. Thus, the possible *OD* pairs of the *OD/QoS* combinations for a particular router module 130 would typically include the *OD* pairs derived from all data packet originations and data packet destinations known to the particular router module 130.

Typically networks 120 and router modules 130 are at times data packet originations and at other times data packet destinations. As such, the same two components of a first *OD* pair could be the components of a second *OD* pair by designating the first pair origination as the second pair destination and the first pair destination as the second pair origination. For small sized data communication systems, a router module 130 stores one potential for each possible *OD/QoS* combination for the entire data communication system. For a larger sized data communication system, an aggregation scheme is used which reduces the number of potentials stored by each router module 130.

An aggregation scheme reduces the number of networks 120 and router modules 130 known to any one router module 130 in order to keep the set of potentials

stored in the router module to a manageable size. As shown in an aggregation hierarchy 510 of Figure 5, this aggregation scheme has a physical portion 512 with physical nodes and a conceptual portion 514 with conceptual nodes. The physical nodes include networks 120 and router modules 130. The conceptual nodes each serve as labels for a particular collection made up of any number of physical nodes and/or other conceptual nodes. The physical nodes of the physical portion 512 are typically found in a particular geographical region. Figure 5 only illustrates a representative example, and the depicted embodiment can have other regions of various sizes and locations.

Although Figure 5 illustrates an aggregation hierarchy in terms of geography, the depicted embodiment is not limited to only hierarchies defined in geographical terms. Other hierarchies defined in terms of an organizational, political, topical, topological, or other types of structures are also included in the present invention. Also, Figure 5 is not intended to limit the present invention to any particular manner or type of geography in defining the physical portion 512 of the aggregation hierarchy 510. The physical nodes can be grouped according to other common geographical regions besides the manner of using cities as regions as illustrated in Figure 5.

For illustration purposes, Figure 5 shows the regions of Seattle, Portland, San Francisco (S.F.), Los Angeles (L.A.), Boston, New York (N.Y.), Atlanta, and Miami having physical nodes 520-534 respectively. The Seattle region and Portland region represented by conceptual nodes 540 and 542 are further aggregated into the Northwest (N.W.) region represented by conceptual node 560. Similarly, conceptual nodes 544 and 546 for the S.F. and L.A. regions respectively are further aggregated into the southwest (S.W.) region represented by conceptual node 562. Conceptual nodes 548 and 550 for the Boston and N.Y. regions respectively are aggregated into the Northeast (N.E.) region represented by conceptual node 564. Conceptual nodes 552 and 554 for the Atlanta and Miami regions respectively are aggregated into the Southeast (S.E.) region represented by conceptual node 566. The conceptual nodes 560 and 562 for the N.W. and S.W. regions respectively are aggregated into the West Coast region represented by conceptual node 570. The conceptual nodes 564 and 566 for the

N.E. and S.E. regions respectively are aggregated into the East Coast region represented by conceptual node 572. The conceptual nodes 570 and 572 for the West Coast and East Coast regions respectively are aggregated into the USA region represented by the conceptual node 574.

Physical nodes of the various regions in the physical portion 512 are directly connected with other physical nodes (typically router modules 130) in other regions of the physical portion. For instance, a physical node 520c in the Seattle region is directly connected to a physical node 522a in the Portland region through link 580. The various regions are connected together via this type of direct connection between physical nodes in one region and other physical nodes (typically router modules 130) in another region. Not all physical nodes in one region need to be connected to another physical node in another region. This is because physical nodes within the same region are connected to one another either directly or through other physical nodes (typically router modules 130) in the same region. Subnetworks formed by the aggregation scheme need to maintain connectivity. Thus, between each pair of nodes in a subnetwork there must be a connection. In the aggregation scheme illustrated in Figure 5, the physical nodes 520a-c in the area of Seattle all share a common conceptual node 540 in the conceptual portion 514 labeled Seattle. Each physical node 520a-c in the area of Seattle stores one potential for each possible *OD/QoS* combination based on the physical nodes 520 composed of networks 120 and router modules 130 in physical area 512 associated with the Seattle conceptual node 540. Also, under the same aggregation scheme, each physical node 520a-c in the Seattle area further stores one potential for each possible *OD/QoS* combination based also on the conceptual nodes that are blackened in on Figure 5. The resultant *OD/QoS* combinations include any *OD* pair combination based on either physical nodes, conceptual nodes or a mixture of both physical and conceptual nodes as the originations and destinations of the *OD* pairs. The conceptual nodes that are blackened in on Figure 5 are Seattle conceptual node 540, Portland conceptual node 542, N.W. conceptual node 560, S.W. conceptual node 564, West Coast conceptual node 570, East Coast conceptual node 572, and USA conceptual node 574. As shown in Figure 5, the physical node 520c is directly connected to

physical node 522a via physical link 580 and physical node 524a via physical link 582. Due to these direct connections, physical node 520c also stores potentials for *OD/QoS* combinations that incorporate one or both of physical nodes 522a and 524a as either origination, destination or both origination and destination. Thus, physical node 520c stores more potentials due to more *OD* pairs than the other physical nodes 520. In an alternate embodiment only physical nodes in the same region as the node in question along with appropriate conceptual nodes are used for possible *OD* pair combinations. For instance in the alternative embodiment, physical node 520c only stores potentials for physical nodes 520a-c in physical portion 512 and the blackened conceptual nodes. In the alternative embodiment, physical node 520c does not store potentials for any other directly connected nodes such as physical nodes 522a and 524a.

There are general rules to designate which physical and conceptual nodes are used to derive *OD* pairs for the *OD/QoS* combinations for a particular physical node. First, a router module 130 (at physical node 520c in the example of Figure 5) stores potentials for each *OD/QoS* combination having an *OD* pair based on the physical nodes (520b and 520c in the example of Figure 5) with which the router module 130 shares a common conceptual node (540 in the example of Figure 5). Second, if the router module (at physical node 520c in the example of Figure 5) is directly connected to other physical nodes (522a and 524a in the example of Figure 5), the other physical nodes are also used to derive other *OD* pairs for *OD/QoS* combinations for which potentials are stored in the router module (at physical node 520c). Finally, the router module (at physical node 520c) will store potentials for *OD/QoS* combinations for *OD* pairs that incorporate one or both of conceptual nodes that are either "lineal ancestors" of the router module (at physical node 520c) or "siblings" of the lineal ancestors. For the example of Figure 5, the conceptual nodes that are lineal ancestors of the physical nodes 520 (including router modules 130) are Seattle conceptual node 540, N.W. conceptual node 560, West Coast conceptual node 570 and USA conceptual node 574. The conceptual nodes in the example of Figure 5 that are siblings of the lineal ancestors are Portland conceptual node 542 (sibling of Seattle conceptual node 540), S.W. conceptual node 562 (sibling of N.W. conceptual

node 560), and East Coast conceptual node 572 (sibling of West Coast conceptual node 570).

The method for optimizing network flow in a data communication system used by the depicted embodiment of the dynamic load balancer unit 216 is further based on a difference function $d(f_{j,ab}, p_{j,ab})$. The difference function $d(f_{j,ab}, p_{j,ab})$ is evaluated for each arc "*ab*" for each *OD/QoS* combination "*j*" for each router module 130 in the data communication system. When the difference function $d(f_{j,ab}, p_{j,ab})$ is evaluated for a particular router module 130, that router module is designated a home router module. The difference function $d(f_{j,ab}, p_{j,ab})$ is evaluated for each arc that either originates from or terminates with the home router module. Router modules 130 that are physically connected to the home router module are designated neighboring router modules. The neighboring router modules share with the home router module 130 the ares either originating or terminating with the home router module potentials. Potentials $p_{j,a}$ and $p_{j,b}$ are at the origination and destination router modules 130 respectively of arc *ab* for *OD/QoS* combination "*j*". Thus, the home router module 130 and neighboring router modules 130 are either origination or destination router modules, depending upon the origination and destination of the particular arcs being evaluated.

For a particular arc "*ab*" involved in the evaluation, the difference function $d(f_{j,ab}, p_{j,ab})$ is evaluated for each common *OD/QoS* combination "*j*" that is known both to the home router module 130 and the neighboring router module 130 of the particular arc "*ab*." The resultant expression for the difference function $d(f_{j,ab}, p_{j,ab})$ for arc "*ab*" and *OD/QoS* combination "*j*" is

$$d\left(f_{j,ab}, p_{j,ab}\right) = \left(r_{j,ab} / 2\right) * \left(\hat{g}_{ab} - f_{j,ab}\right)^2 - f_{j,ab} * \left(p_{j,a} - p_{j,b}\right)$$

where

$$\hat{g}_{ab} = \alpha g_{j,ab} - \beta\left(g_{ab} - \sum_{k \neq j} f_{k,ab}\right)$$

The potentials $p_{j,a}$ and $p_{j,b}$ are potentials for router modules 130 "a" and "b" respectively for the "j" *OD/QoS* combination. Flows $f_{k,ab}$ where $k \neq j$ are flows for the arc "ab" associated with *OD/QoS* combinations other than *OD/QoS* combination "j".

The first term, $(r_{j,ab}/2)*(\hat{g}_{ab} - f_{j,ab})^2$, of the difference function $d(f_{j,ab}, p_{j,ab})$ is a simplified version of the expression for $E_{\alpha,\beta}(f_{j,ab})$ where only one arc "ab" and one *OD/QoS* combination "j" is considered. The difference function $d(f_{j,ab}, p_{j,ab})$ is subject to the following boundary conditions:

$$\hat{l}_{j,ab} \leq f_{j,ab} \leq \hat{u}_{j,ab}$$

where

$$l_{ab} \leq \sum_j f_{j,ab} \leq u_{ab},$$

$$\hat{l}_{j,ab} = \max\left\{l_{j,ab}, l_{ab} - \sum_{k \neq j} f_{k,ab}\right\},$$

$$\hat{u}_{j,ab} = \min\left\{u_{j,ab}, u_{ab} - \sum_{k \neq j} f_{k,ab}\right\},$$

and where $l_{j,ab}$ and l_{ab} are lower limits, and $u_{j,ab}$ and u_{ab} are upper limits on $f_{j,ab}$ and f_{ab}, *respectively*.

An intuitive explanation of the difference function $d(f_{j,ab}, p_{j,ab})$ is that the first term related to $E_{\alpha,\beta}(f_{j,ab})$ represents undesirable results and conditions of data packet traffic being transmitted over a data communication system, such as delays, poor utilization of resources, poor adaptability to sudden changes, etc. For a given arc, when the goal parameter, $\hat{g}_{ab}$, equals zero and the resistance parameter equals the inverse of capacity of the arc, then the first term of the difference function $d(f_{j,ab}, p_{j,ab})$ equals the number of jobs per unit time that have to go into a waiting queue because a router module 130 is busy. The second term $f_{j,ab}*(p_{j,a} - p_{j,b})$ represents effective utilization of

the data communication system as subject to constraints imposed by the data communication system such as flow constraints. The difference function $d(f_{j,ab},p_{j,ab})$ thus represents a difference between desirable and undesirable aspects involved with the data communication system.

As part of finding a local equilibrium point, by differentiating the difference function $d(f_{j,ab},p_{j,ab})$ at each arc between router modules with respect to the flow $f_{j,ab}$ of the *OD/QoS* combination "*j*" and arc "*ab*", setting the resultant differential equal to zero, and solving for the flow, $f_{j,ab}$, a response function results for flow $f_{j,ab}$. As illustrated in Figure 6, for a particular arc "*ab*" 610 having beginning router module "*a*" 620 and ending router module "*b*" 630 and for a common *OD/QoS* combination "*j*", the response function of the optimal flow $f_{j,ab}$ 640 is:

$$f_{j,ab}(p) = \max\left\{\hat{l}_{j,ab}, \min\left\{\hat{u}_{j,ab}, \hat{g}_{ab} - \left(p_{j,a} - p_{j,b}\right)/r_{j,ab}\right\}\right\}$$

where min and max are minimum and maximum functions which choose the minimum or maximum term of the pair of terms that are bracketed. For instance, max (*a*,*b*) chooses the maximum value between *a* and *b*. Thus, max (*a*, min (*b*,*c*)) first chooses the minimum between *b* and *c* and then chooses the maximum between *a* and the minimum choice between *b* and *c*. Minimum flow for $f_{j,ab}$ is typically zero. Maximum flow for $f_{j,ab}$ is typically a percentage of arc capacity.

The goal parameter, $\hat{g}_{ab}$, and resistance parameter, $r_{j,ab}$, are for the particular arc "*ab*" and *OD/QoS* combination "*j*". The potential $p_{j,a}$ is for the arc beginning router module "*a*" 620 and the potential $p_{j,b}$ is for the arc ending router module "*b*" 630. Both potentials $p_{j,a}$ and $p_{j,b}$ are associated with a common *OD/QoS* combination "*j*". Typically, the origination and destination nodes for an *OD/QoS* combination are not the arc beginning and arc ending router modules 620 and 630 respectively. Also, the origination and destination nodes for an *OD/QoS* combination can be a combination of physical and/or conceptual nodes whereas the arc beginning and arc ending nodes are always physical since they are router modules 130.

Summing these optimal flows for arc beginning "*a*" and arc ending "*b*" router modules 620 and 630 for arc "*ab*" over the *OD/QoS* combinations that are known

to the arc beginning and ending router modules results in the optimal flow for arc "*ab*". This response function of flow describes an optimized flow based on certain values for potentials. These potential values further satisfy a constraint based on how data packets are created and terminated in the data communication system.

In the depicted embodiment, the potentials are independent variables and the flows are dependent variables. However, in other embodiments, different subsets of flows and potentials are used so that some flows and potentials are dependent variables and other flows and potentials are independent variables.

The depicted embodiment of the invention is configured so that all data packet traffic originates and terminates at network terminals of network 120 or router modules 130. This means that all data packet traffic having a common destination, either created at a particular router module 130 or arriving at the particular router module, must equal the data packet traffic with the common destination leaving the router module 130. This requirement is referred to as a conservation of flows or a flow constraint. When implementing the method of the depicted embodiment, the router modules 130 scattered throughout a data communication system frequently adjust the data packet routes in the data communication system to pursue the goal of maximizing net benefits for stochastic distributions of network loads and stochastic changes in topology. The adjustments are done by manipulating potential values according to the conservation of flows.

Adjustments are first needed when a data communication system 110 is initially brought into service. Adjustments are subsequently needed due to the continually changing demands or loads by users of the data communication system and changes in the topology of the data communication system due to growth of the data communication system or equipment failures within the existing data communication system. The method of the depicted embodiment incorporated into the dynamic load balancer units 216 of the router modules 130 is responsible for these adjustments made by the router modules.

As stated above, the dynamic load balancer unit 216 uses information collected by the neighborhood supervisor unit 214. This collected information can now

be described in terms of the response function for flow, $f_{j,ab}$, as discussed above. The discussion will be focused on a particular router module 130 designated the "home" router module, however, the home router module discussed is representative of all router modules in that all router modules also perform the steps described.

In general, there are two main approaches that are used in the present embodiment to determine the flows for a given home router module 130. In both approaches, arcs are distinguished as being either supply arcs or non-supply arcs. A supply arc either directly connects an origination network 120 or origination user device of a network for a given *OD/QoS* combination to the given home router module 130 or a router module between the network or user device and the home router mode where there is no branching of the arcs for the home router module and the network or user device. Non-supply arcs are also directly connected to the given home router module but are not directly connected to an origination node.

For both approaches, the goal parameters for a supply arc for a given *OD/QoS* combination are set equal to the user demand or load placed on the supply arc. For instance, if user demand or load was 100 Mbits/sec on a supply arc for a given *OD/QoS* combination then the goal parameter for the same supply arc for the same *OD/QoS* combination would be set to 100 Mbits/sec. The neighborhood supervisor unit 214 senses what the demand or load is from any networks 120 on network devices that are directly connected to the home router module 130. Also, for both approaches the resistance parameter for a given *OD/QoS* combination would be set to a relatively large number for a supply arc. The large number means that there is a high cost incurred if the goal of the goal parameter (in this case 100 Mbits/sec) is not met. For a non-supply arc the resistance parameter for a given *OD/QoS* combination is set to the inverse of the arc capacity for the given *OD/QoS* combination.

The two approaches to solve for flows for a given router module differ in how they deal with the goal parameters for non-supply arcs. The first approach solves for flow for one *OD/QoS* combination at a time whereas the second approach solves for flow for all *OD/QoS* combinations at once. For the first approach, the goal parameter for a non-supply arc is set to the negative of the historical flow for an *OD/QoS*

combination for the arc. The historical flow for an arc is the most recent flow amount recorded. Often times historical flows tend to change more slowly than demand or load, especially for arcs that have become a major artery for a data communication system. For instance, if the flow on a non-supply arc was most recently recorded as 50 Mbits/sec for an *OD/QoS* combination, then the goal parameter for that arc would be -50Mbits/sec.

With the first approach, for a given home router module 130, the dynamic data flow splitter unit 220 determines historical flows from header information on data packets sent from the home router module. These historical flows of the home router module for sent packets are stored in the dynamic data flow splitter unit 220. The historical flows are represented by averages, variances and other expressions known in the art of statistics and mathematics. The neighborhood supervisor unit 214 queries the dynamic data flow splitter unit 220 for the historical flows of sent data packets as needed for flow calculations or to send to other neighboring router modules. The historical flows of sent data packets are for arcs directly connected to the given home router module 130 handling packets sent by the home router module to networks or other router modules. In the depicted embodiment, the neighborhood supervisor unit 214 of the given home router module 130 also receives updated historical flow information from other neighborhood supervisors for arcs that are directly connected to the given home router module and carry data packets to the given home router module. A special case of the first approach is where all historical flows are equal to zero. This special case occurs at such times as the initial start-up phase of a data communication system.

The second approach to solve for flows for a given home router module 130 uses computed flows instead of historical flows. Flows are determined for all *OD/QoS* combinations at once through an iterative process. For a K^{th} iteration, the overall flow for an arc is determined by summing all flows for the *OD/QoS* combinations for the arc solved during the K^{th} iteration. The goal parameter for the non-supply arcs for the $K+1^{th}$ iteration are then equal to the negative of the overall flows for the arcs for the K^{th} iteration.

The techniques used to solve in this iterative fashion the flows for a home router module 130 for all *OD/QoS* combinations at once require extensive computing capability. Some embodiments use the second approach to solve for flows for all *OD/QoS* combinations at once. The depicted embodiment as incorporated into the methods of the described neighborhood supervisor unit 214 uses the first approach to solve for flows one *OD/QoS* combination at a time for a given home router module starting with Step 710 of Figure 7.

With the first approach, the neighborhood supervisor unit 214 of the home router module 130 collects potentials and historical flows from all neighboring router modules 130 directly connected to the home router module in step 720 of the flowchart of Figure 7 showing the method used by the depicted embodiment. The collected potentials and historical flows are for common *OD/QoS* combinations known to both the home router module and at least one neighboring router module. Either the neighboring router modules or the home router module initiates transmission and reception of the potentials and historical flows collected from the neighboring router modules to the home router module. In the depicted embodiment, the neighboring router modules initiate the transmission of their potentials and historical flows to the home router module. However, in other embodiments, the home router module initiates the transmission. The dynamic load balancer unit 216 of the home router module 130 then uses the collected potentials and historical flows along with the potentials and historical flows of the home router module to calculate, via the flow response function, the data packet flows for each *OD/QoS* combination to and from the home router module for all arcs connected to the home router module. For each connection between the home router module 130 and neighboring router modules or networks 120, there are two arcs as described above. One arc has the home router module as the origination node and the other arc as the destination node. In other embodiments, only one-sided connections are possible so that for each connection there is only one arc.

The neighborhood supervisor unit 214 then uses the historical flows and demand or load flows to update the goal parameters for all the non-supply and supply arcs respectively of the home router module 130 in step 722 of Figure 7. Typically, the

goal parameters for non-supply arcs are set to the negative of the historical flow for a given *OD/QoS* combination and the goal parameters for the supply arcs are set to the demand or load flow on the arc. The neighborhood supervisor unit 214 also identifies any arcs that have zero overall flow based on their historical or demand/load flow so that any subsequent modifications of the home potentials are done with the zero flow conditions in mind as explained further below.

Initially, no *OD/QoS* combinations are selected by the neighborhood supervisor unit 214, so the outcome of decision 724 of Figure 7 regarding whether all *OD/QoS* combinations have been selected is "no." In such case, the neighborhood supervisor unit 214 selects an *OD/QoS* combination in step 726 that has not been chosen since the last time the start step 710 was performed. For all arcs either originating from or terminating into the home router module, the neighborhood supervisor unit 214 of the home router module passes, in step 728, all the potentials of the home router module and those collected from neighboring router modules for the selected *OD/QoS* combination to the dynamic load balancer unit 216. The neighborhood supervisor unit 214 also passes all the goal parameters and resistance parameters for the arcs of the particular router module 130 based on the most recent historical and demand/load flows collected by the neighborhood supervisor unit 214 to the dynamic load balancer unit 216.

The dynamic load balancer unit 216 then determines, for the selected *OD/QoS* combination, the flows across the arcs of the home router module. The dynamic load balancer unit 216 then determines, in decision step 730 of Figure 7 whether the determined flows are conserved for the selected *OD/QoS* combination or whether the time period for iteration has expired. Flows are conserved when the incoming flows plus flows originated by the home router module equals the flows leaving the home router module. If flows are not conserved for the selected *OD/QoS* combination or the time for iteration has not expired, the decision step in 730 is "no" and the method branches to step 732 where the dynamic load balancer unit 216 adjusts the potentials for the home router module 130 for the selected *OD/QoS* combination. After the home potentials are adjusted in step 732, flow conservation is again checked

in decision step 730. The home potentials are iteratively adjusted through the combination of potential adjustment in step 732 and flow conservation verification in decision 730 until flow is conserved and decision step 730 is "yes" and the method branches back to step 722 to update the goal parameters. Iteration techniques to allow for rapid convergence to solutions involving iteration are known in the art and include over or under-relaxation techniques. In one embodiment, the potentials are adjusted one at a time by the iteration process for a given *OD/QoS* combination until flow is conserved. In another embodiment all the potentials for a router module for all the *OD/QoS* combinations associated with the router module are adjusted for each iteration of a Gaussian or other second order method of determining solutions by block techniques known in the art.

If the neighborhood supervisor unit 214 has identified any zero flow arcs that have zero overall flow when the goal parameters are initially updated in step 722 immediately after the historical and demand flows are obtained in step 720, then during the home potential adjustment in step 732, a goal parameter associated with a zero flow arc can be set to the negative of any calculated flow for the zero flow arc resulting from adjustments of potentials for an *OD/QoS* combination to ensure that the flow for the zero flow arc remains small. The new value for the goal parameter for the zero flow arc is then used to update the goal parameters in step 722. At the completion of adjusting potentials for all *OD/QoS* combinations, the updated goal parameter for a zero flow arc will then be the summation of flows over all *OD/QoS* combinations for the particular zero flow arc due to adjustment of potentials.

Once any goal parameters are updated in step 722 another *OD/QoS* combination is selected which had not been previously selected since the last performance of start step 710 and steps 728-732 are again performed for the newly selected *OD/QoS* combination. After steps 728-732 are completed, again any changed goal parameters are updated. The process similarly cycles through the *OD/QoS* combinations until decision step 724 determines that all *OD/QoS* combination have been selected, whereby the "yes" branch is taken and decision step 734 determines if any potentials of the home router module 130 had been adjusted by step 732.

If any home potentials were adjusted, then decision step 734 branches to step 736 of Figure 7 where the neighborhood supervisor unit 214 of the home router module 130 reports the adjusted home potentials and historical flows to the neighboring router modules. Decision step 738 then determines if the period of time since the neighborhood supervisor unit 214 of the home router module 130 last reported home potentials to neighboring router modules is greater than a predetermined amount of time, T_R. If the period is not greater then T_R, decision step 738 branches back to the start step 710. If the period is greater than T_R, decision step 738 branches to decision step 740, the same step to which decision step 734 branches if no home potentials were adjusted. In decision step 740 a determination is made as to whether any adjusted home potentials have changed more than a predetermined threshold T_T since the last time electronic tables of the dynamic load balancer unit 216 storing the home potentials have been updated with adjusted home potentials. If any changes in the adjusted home potentials are greater than T_T, decision step 740 branches to step 742 wherein the electronic tables are updated with the latest values for the home potentials. The dynamic load balancer unit 216 subsequently updates the routing table unit 218. If there are no changes in the adjusted home potentials greater than T_T, decision step 740 branches to the start step 710.

The present invention is not limited to the particular routines used by the depicted embodiment to report updated potentials to neighboring router modules 130 or in updating electronic tables storing home potentials. Other embodiments use other methods, intervals, or thresholds to determine when to report to neighboring router modules 130 and when to update electronic tables storing home potentials.

In an alternative embodiment, the resistance parameter for each particular arc is a function of the flows of each *OD/QoS* combination. For instance, for a particular *OD/QoS* combination, the resistance, r, is as follows:

$$r = \frac{d^2}{df^2}\left(f/(c-f)\right)$$

where f is the flow for an arc for an *OD/QoS* combination and c is the capacity or residual capacity of the arc. It follows that since the response function expresses the

dependency of flow for an *OD/QoS* combination for an arc, the resistance parameters are also expressible in terms of potentials.

From the foregoing it will be appreciated that, although specific embodiments of the invention have been described herein for purposes of illustration, various modifications may be made without deviating from the spirit and scope of the invention. Accordingly, the invention is not limited except as by the appended claims.

CLAIMS

1. In a data communication system having network traffic with randomly distributed demands and flows, a network traffic director system comprising:

router modules configured to direct data packets in the data communication system, each router module being a home router module comprising:

a neighborhood supervisor configured to send home potentials of the home router module to neighborhood supervisors of neighboring router modules and to receive neighbor potentials of the neighboring router modules from neighborhood supervisors of neighboring router modules;

a dynamic load balancer configured to determine flows based on the home and neighbor potentials, to adjust the home potentials if first conditions including flow conditions are not met, and to update routing tables if second conditions based on the adjusted home potentials are met; and

a dynamic data flow splitter configured to receive data packets from networks and router modules, to select a portion of the data communication system for each data packet received based on the updated routing tables wherein each received data packet is transmitted to the portion of the data communication system selected by the dynamic data flow splitter for the received data packet.

2. The network traffic director system of claim 1 wherein each of the home potentials of each home router module is associated with a pair of nodes of the data communication system comprising an origination node and a destination node.

3. The network traffic director system of claim 1 wherein each of the home potentials of each home router module is associated with a quality of service level.

4. The network traffic director system of claim 1 wherein the neighborhood supervisor is configured to send the home potentials to neighborhood supervisors of neighboring router modules when the flow conditions are not met.

5. The network traffic director system of claim 1 wherein the flow conditions comprise conservation of flows.

6. The network traffic director system of claim 1 wherein the dynamic load balancer is configured to determine flows by using a response function associated with optimizing at least one of a penalty function or a merit function for stochastic demands and flows of the data communication system.

7. The network traffic director system of claim 1 wherein the dynamic data flow splitter is configured to select the portion of the data communication system for each data packet received further based on a Markov protocol.

8. The network traffic director system of claim 1 wherein the dynamic data flow splitter is configured to select the portion of the data communication system for each data packet received further based on a Routing Wheel protocol.

9. The network traffic director system of claim 1 wherein the dynamic data flow splitter is further configured to select a portion of the data communication system based on previously selected portions of the data communication system when an address for a received data packet is a same type as an address stored in the dynamic data flow splitter.

10. The network traffic director system of claim 1 wherein the dynamic data flow splitter is further configured to select of a portion of the data communication system based on a quality of service level associated with a data packet.

11. In a communication system having network traffic with flows, a network traffic director system comprising:

router modules, each router module being a home router module configured to store and adjust home potentials of the home router module, and configured to receive and

store neighbor potentials of neighboring router modules, the home router modules configured to determine ideal data flows using the home and neighbor potentials with an optimization of at least one of a merit function or a penalty function involving stochastic changes in at least one of demands or topology in the communication system, the home router modules configured to receive and route network traffic based on the home and neighbor potentials.

12. The network traffic director system of claim 11 wherein home router modules are configured to be a neighbor router module of other home router modules.

13. The network traffic director system of claim 11 wherein the router modules are configured to continue adjusting home values to seek flow conservation until flow conservation is satisfied or until a time period has expired.

14. The network traffic director system of claim 11 wherein the merit or penalty function are at least locally approximated by a quadratic function of the flows.

15. The network traffic director system of claim 11 wherein the neighbor potentials are stored according to an aggregation scheme.

16. The network traffic director system of claim 11 wherein the ideal data flows are further based on resistance of arcs associated with the router module wherein each resistance of an arc is configured to be a function of arc capacity.

17. The network traffic director system of claim 11 wherein the ideal data flows are further based on resistance of arcs associated with the router module wherein each resistance of an arc is configured to be a function of flows of the arc.

18. The network traffic director system of claim 11 wherein the neighbor router modules are connected to the home router modules without any other router modules therebetween.

19. The network traffic director system of claim 11 wherein the neighbor router modules are configured to be associated with the home router module through at least one of a political, organizational, topological, topical, or geographical relation.

20. The network traffic director system of claim 11 wherein each router module is configured to adjust home potentials to an equilibrium point for a difference function associated with flows of arcs associated with the router module, the difference function associated with differences between home potentials and neighbor potentials.

21. The network traffic director system of claim 11 wherein each router module stores adjusted home values when adjustment of home values is greater than a threshold.

22. The network traffic director system of claim 11 wherein each home router module transmits adjusted home values to its neighbor router modules based on how much time has elapsed since last transmitting adjusted home values.

23. The network traffic director system of claim 11 wherein the merit function or penalty function involves costs of congestion comprising at least one of traffic delays, high latency, diminished throughput, lost packets, or unresponsiveness to sudden changes in topology or loading.

24. The network traffic director system of claim 11 wherein the neighbor router modules are connected to the home router modules wherein each connection has only one arc.

25. In a data communication system having network traffic, a network traffic director method comprising:

storing home potentials associated with a home router module;

receiving and storing neighbor potentials associated with neighbor router modules of the home router module;

determining ideal flows based on the home and neighbor potentials and optimizing at least one of a merit function or a penalty function for random changes in demands or topology in the data communication system;

determining conservation of flows for the home router module;

adjusting the home potentials to approach a state where flows are conserved if determined not to be conserved;

reporting adjusted home potentials to the neighbor router modules; and

routing network traffic based on home and neighbor values.

26. The network traffic director method of claim 25 wherein the routing further includes using at least one of a Markov protocol, a Routing Wheel protocol, stored data packet addresses, or a level of quality of service of received communication traffic to be routed.

27. The network traffic director method of claim 25 wherein the determining ideal flows further involves a difference function associated with flows, home potentials, and neighbor potentials.

28. The network traffic director method of claim 25 wherein the merit or penalty functions are at least locally approximated by a quadratic function of flows.

29. In a data communication system having network traffic, a network traffic director system comprising:

means for storing home potentials associated with a home router module;

means for receiving and storing neighbor potentials associated with neighbor router modules of the home router module;

means for determining ideal flows based on the home and neighbor potentials and optimizing at least one of a merit function or a penalty function for random changes in demands or topology in the data communication system;

means for determining conservation of flows for the home router module;

means for adjusting the home potentials to approach a state where flows are conserved if determined not to be conserved;

means for reporting adjusted home potentials to the neighbor router modules; and

means for routing network traffic based on home and neighbor values.

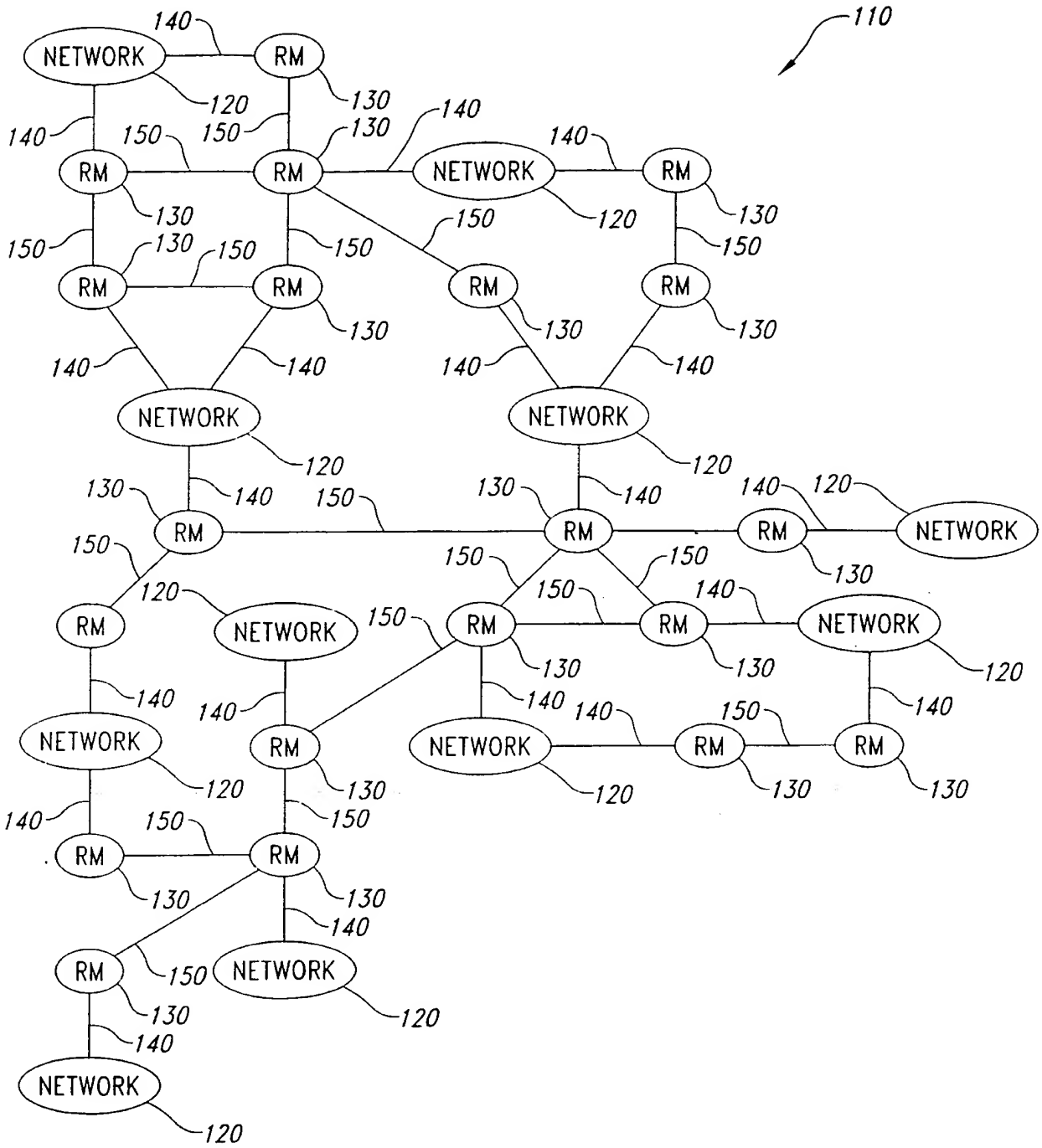


Fig. 1

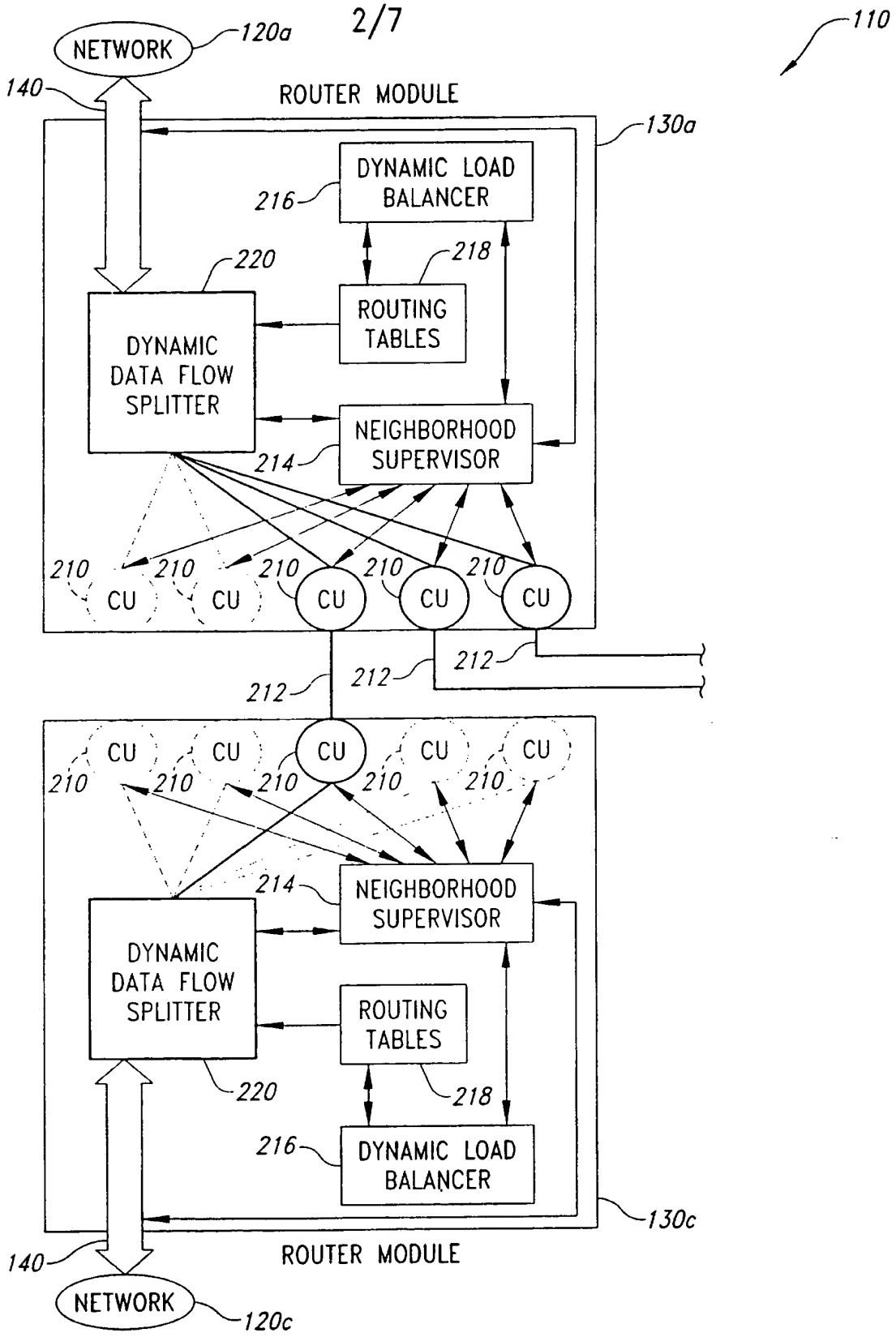


Fig. 2A

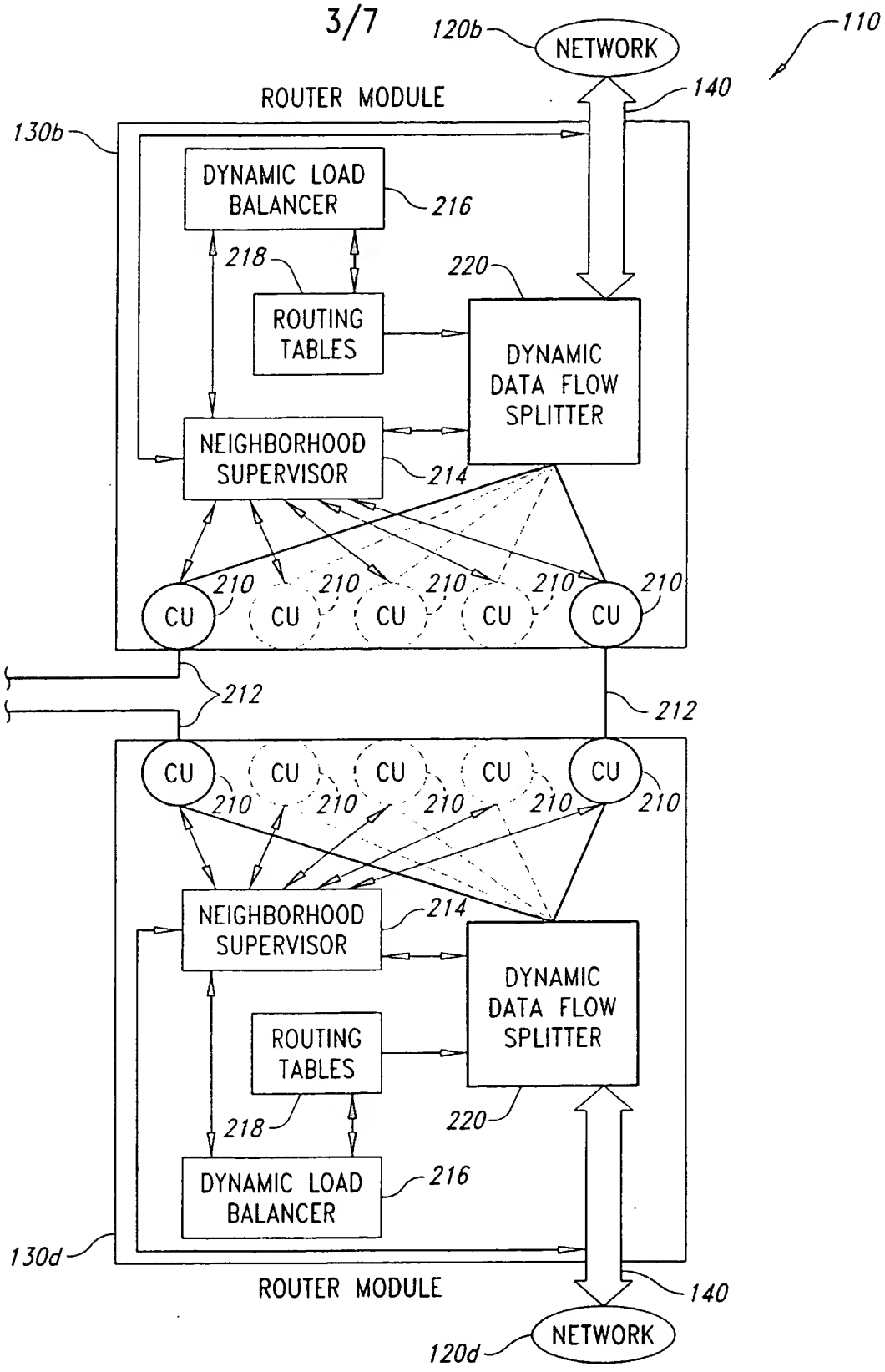


Fig. 2B

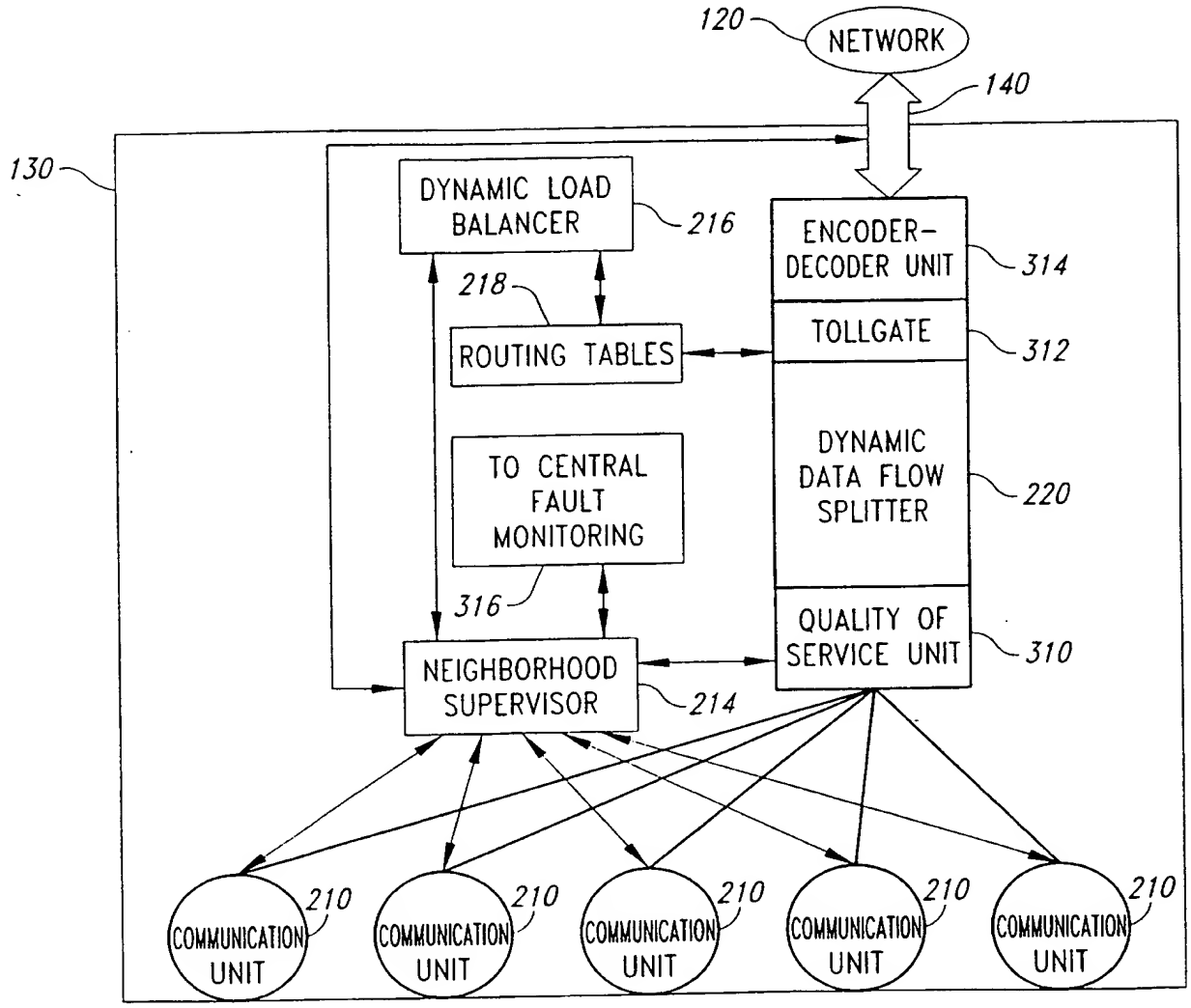


Fig. 3

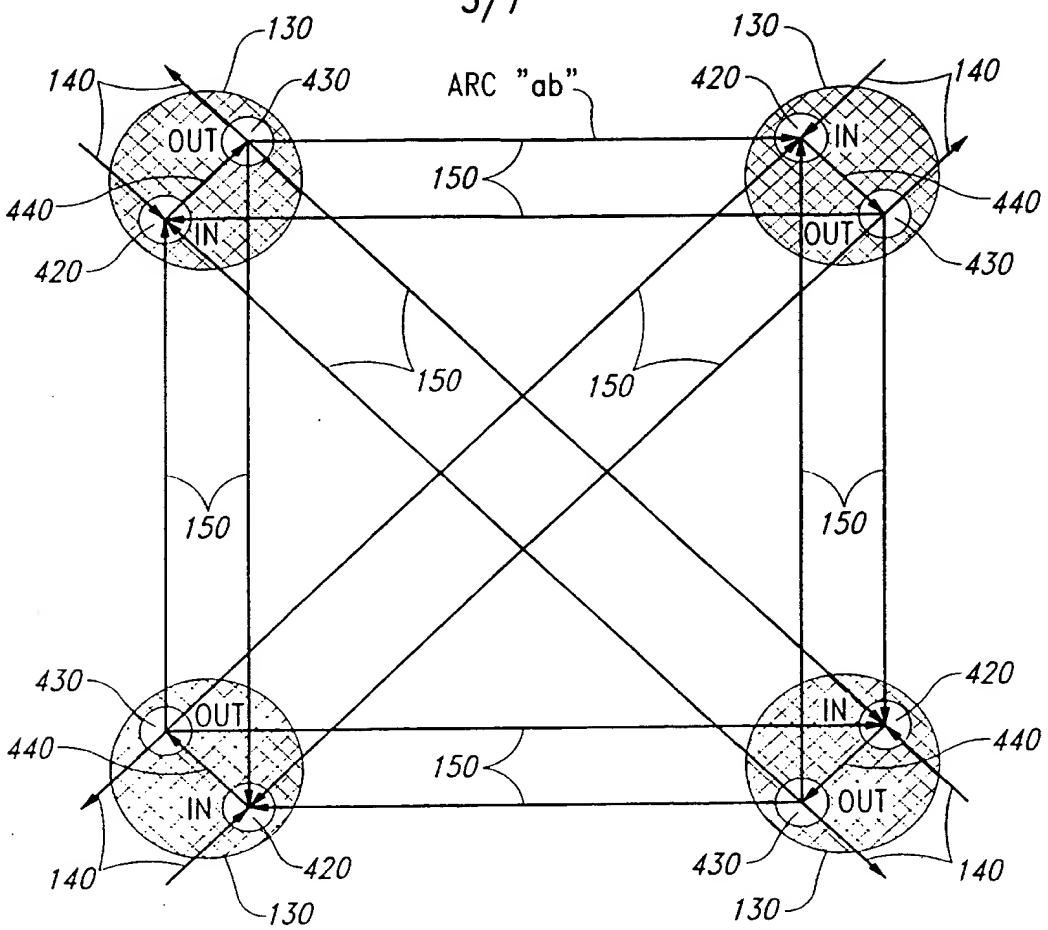

130
140
430
ARC "ab"
420
130
140
OUT
IN
150
440
440
IN
OUT
420
430
150
150
150
150
430
420
OUT
IN
440
440
150
IN
OUT
140
420
430
140
130
130

Fig. 4

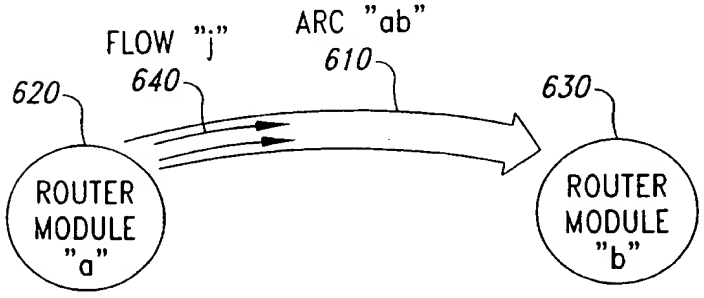

FLOW "j"
ARC "ab"
610
640
620
630
ROUTER MODULE "a"
ROUTER MODULE "b"

Fig. 6

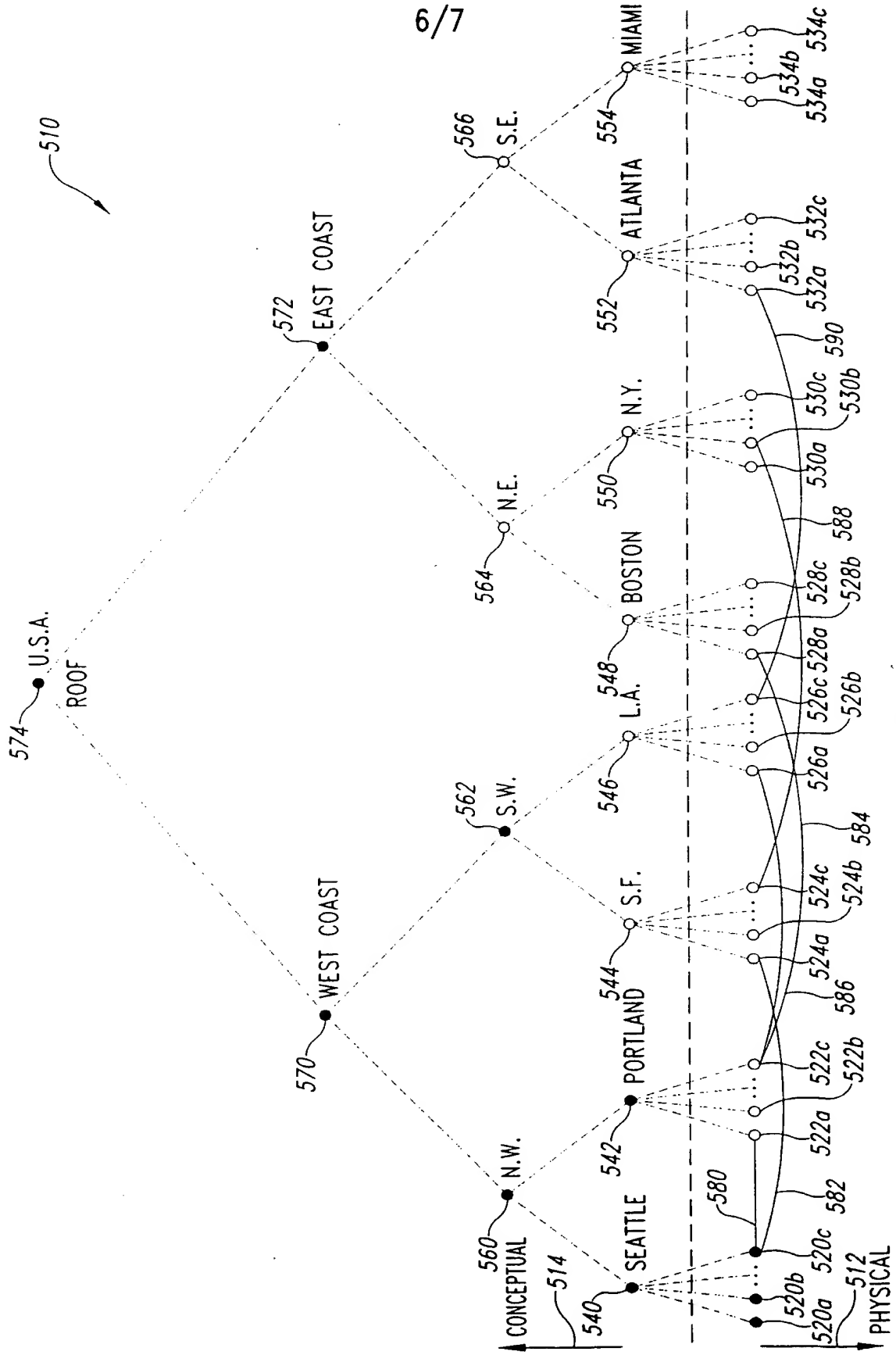

510
U.S.A.
ROOF
574
EAST COAST
572
WEST COAST
570
N.E.
564
S.E.
566
S.W.
562
N.W.
560
BOSTON
548
NEW YORK
550
ATLANTA
552
MIAMI
554
L.A.
546
S.F.
544
PORTLAND
542
SEATTLE
540
CONCEPTUAL
514
PHYSICAL
512
520a
520b
520c
522a
522b
522c
524a
524b
524c
526a
526b
526c
528a
528b
528c
530a
530b
530c
532a
532b
532c
534a
534b
534c
580
582
584
586
588
590

Fig. 5

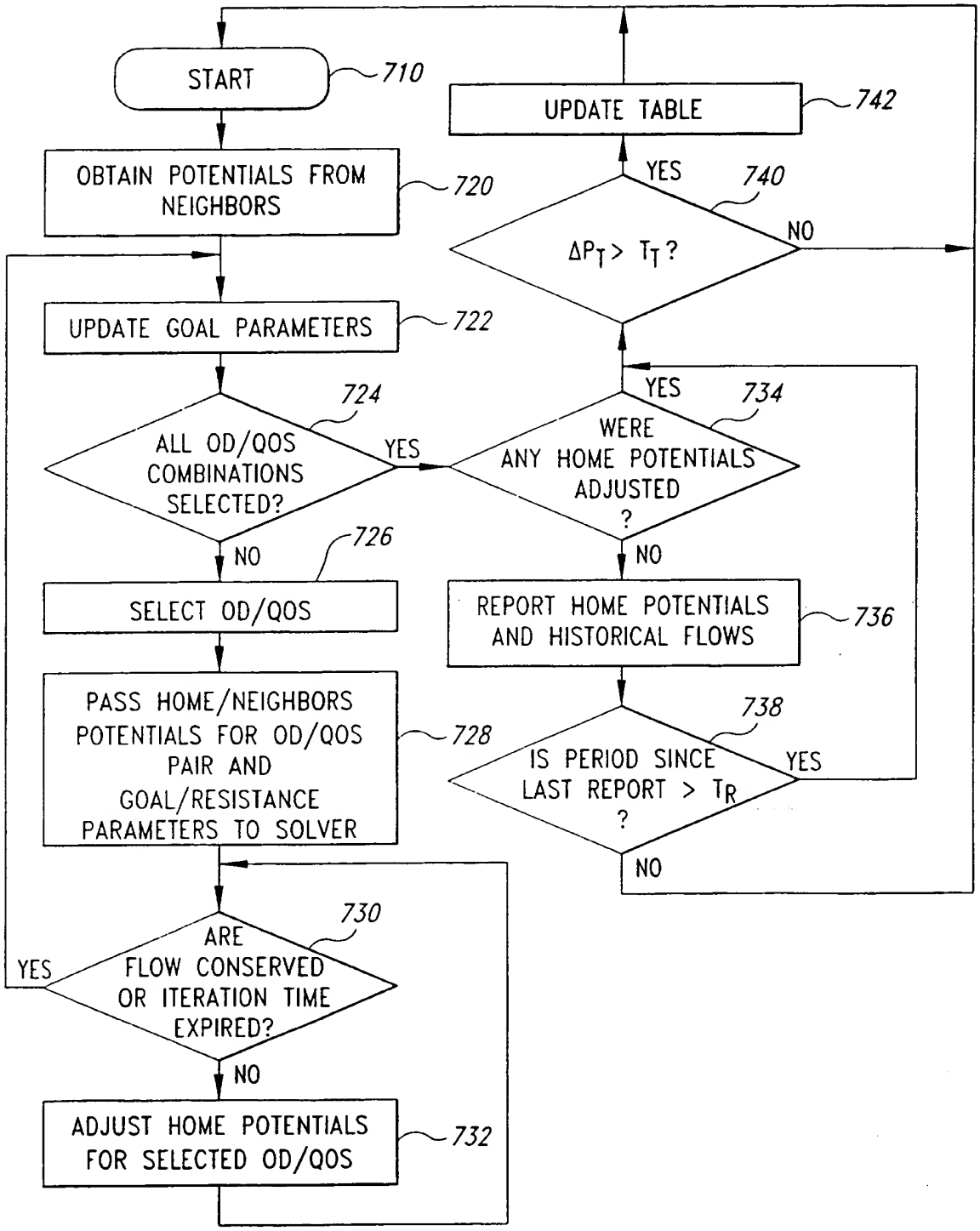


Fig. 7

INTERNATIONAL SEARCH REPORT

Inter mal Application No
PCT/US 99/24145

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 H04L12/56

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 H04L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	NOBUYUKI OBA ET AL: "AN ADAPTIVE NETWORK ROUTING METHOD BY ELECTRICAL-CIRCUIT MODELING" PROCEEDINGS OF THE ANNUAL JOINT CONFERENCE OF THE COMPUTER AND COMMUNICATIONS SOCIETIES (INFOCOM),US,LOS ALAMITOS, IEEE COMP. SOC. PRESS, vol. CONF. 12, 1993, pages 586-592, XP000399038 ISBN: 0-8186-3580-0 the whole document	1,11,25, 29

-/—

[X] Further documents are listed in the continuation of box C.

[ ] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 March 2000	23/03/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax (+31–70) 340–3016	RAMIREZ DE AREL.., F

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Inte. onal Application No
PCT/US 99/24145

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	VAZAO T M ET AL: "A DYNAMIC ROUTING SCHEME FOR CONNECTIONLESS NETWORKING IN B-ISDN" PROCEEDINGS OF INTERNATIONAL CONFERENCE ON COMPUTER COMMUNICATION,ÄS.L.Ü: ÄS.NÜ, vol. CONF. 13, 1997, pages 117-125, XP000753887 ISBN: 2-7261-1104-1 page 118, right-hand column, paragraph 4 -page 120, left-hand column, paragraph 2	1
A	KRISHNAN R ET AL: "AN APPROACH TO PATH-SPLITTING IN MULTIPATH NETWORKS" PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON COMMUNICATIONS (ICC),US,NEW YORK, IEEE, vol. -, 1993, pages 1353-1357, XP000448363 ISBN: 0-7803-0950-2 page 1353, right-hand column, paragraph 3 -page 1356, right-hand column, paragraph 1	1

1